TECOLOTE RESEARCH, INC.

Bridging Engineering and Economics

Since 1973

Prediction Interval Analysis for Nonlinear Equations

SCEA National Conference

15 June 2006

Dr. Shu-Ping Hu

■ Los Angeles ■ Washington, D.C. ■ Boston ■ Chantilly ■ Huntsville ■ Dayton ■ Santa Barbara
■ Albuquerque ■ Colorado Springs ■ Columbus ■ Ft. Meade ■ Ft. Monmouth ■ Montgomery ■ Ogden
■ Patuxent River ■ Pensacola ■ San Diego
■ Charleston ■ Cleveland ■ Denver ■ New Orleans ■ Oklahoma City ■ Silver Spring ■ Warner Robins AFB ■ Vandenberg AFB

TECOLOTE RESEARCH, INC.

Objectives

- **We need to generate prediction intervals (PI) for nonlinear equations to perform cost risk analysis**
  - CO$TAT currently generates confidence intervals for the estimated coefficients in the nonlinear regression output, but does not provide prediction intervals for the estimated values
  - Unlike OLS, no convenient way to construct PIs for nonlinear CERs (or for equations evaluated using the MUPE option)
- **Options:**
  - Evaluate the Bootstrap method to determine the effort required to implement it in CO$TAT and if it is an appropriate solution
  - Investigate whether an analytic method is readily available
    - Find a good approximation for the variance of predicted y (i.e., $\hat{y}$)
    - Derive an alternative analytic method to compute a reasonable PI for any nonlinear CER

Outline

- **Background**
- **Possible downside of using the Bootstrap method**
  - Explored Bootstrap related comments/results from Reference 1
- **Literature search on PIs for nonlinear equations, including MUPE CERs**
- **Derive solution using Taylor series expansion**
- **Develop PIs in Excel for five sample cases**
- **Compare solutions with SAS and CO$TAT**
- **Conclusions**

1. Book, S., "Prediction Intervals for CER-Based Estimates (With Cost Driver Values Outside the Data Range)," 37th DoDCAS, Williamsburg, VA, 10-13 February 2004

TECOLOTE RESEARCH, INC.

Background

- **Nonparametric analysis has a wide use in applied statistics**
- **MCR suggested using the Bootstrap method to construct Prediction Intervals (with no distribution assumption) for general nonlinear equations, including MUPE CERs**

ote: Slides with MCR headings are from Reference 2

1. Book, S., “Prediction Intervals for CER-Based Estimates (With Cost Driver Values Outside the Data Range),” 37th DoDCAS, Williamsburg, VA, 10-13 February 2004
2. Book, S., “Prediction Bounds for General-Error-Regression CERs,” 39th DoDCAS, Williamsburg, VA, 14-17 February 2006

Downside of Using the Bootstrap - General

- **The actual implementation can be extremely tedious.**
  - At least 1000 Bootstrap samples are needed in many applications
  - The Bootstrap method has great potential in many areas, but its computer intensive nature will continue to hinder its use
- **For nonlinear CERs, the solution might not be achievable due to fitting hundreds of nonlinear equations, which may not converge.**
- **Results may not be reliable for small samples.**
  - The Bootstrap can work well for large samples, but may not be reliable for small number of data points (say, 5, 10, or even 20), regardless of how many Bootstrap samples (iterations) are used

TECOLOTE RESEARCH, INC.

MCR: Prepare Database for Bootstrap Sampling – Calculate the Residuals

Calculate the Residuals from an OLS regression result:

x Values (Cost Driver)	y Values (Actual Costs)	Predicted y Values (Cost Estimates)	Residuals = Actuals-Estimates
7.9	3.595	3.699	0.104
8.2	1.900	4.005	2.105
9.8	3.300	5.635	2.335
11.5	10.900	7.367	-3.533
16.4	15.434	12.358	-3.076
19.7	16.074	15.720	-0.354
23.6	17.274	19.693	2.419

Note: CER derived from x and actual y values is $y = a + bx$, where a = -4.348 and b = 1.0187.

MAD of % errors is 36%, CoV = 28%, R^2 = 86%, Adj.R^2 = 83% & SEE = 2.75

Note: Sum of the residuals is **zero**.

MCR: Draw Random Samples of Residuals

Residuals are drawn randomly from the residual column:

x Values (Cost Driver)	Residual Samples:	#1	#2	#3	#4	#5	#6	...
7.9	1st Residual	0.104	2.105	2.335	-3.533	-3.076	2.419	...
8.2	2nd Residual	-0.354	2.419	0.104	-0.354	2.335	2.419	...
9.8	3rd Residual	0.104	-3.533	-3.533	2.105	2.335	-0.354	...
11.5	4th Residual	2.419	-3.533	2.419	2.419	2.105	2.419	...
16.4	5th Residual	2.105	2.105	-3.533	2.105	-0.354	-0.354	...
19.7	6th Residual	2.105	-3.533	2.105	2.105	2.419	-3.076	...
23.6	7th Residual	-3.533	2.419	-0.354	0.104	2.419	0.104	...

Note: Sampling is done "with replacement", so some residuals will appear more than once in the same sample.

TECOLOTE RESEARCH, INC.

MCR: Compute Bootstrap Samples of "Possible" Actual *y* Values

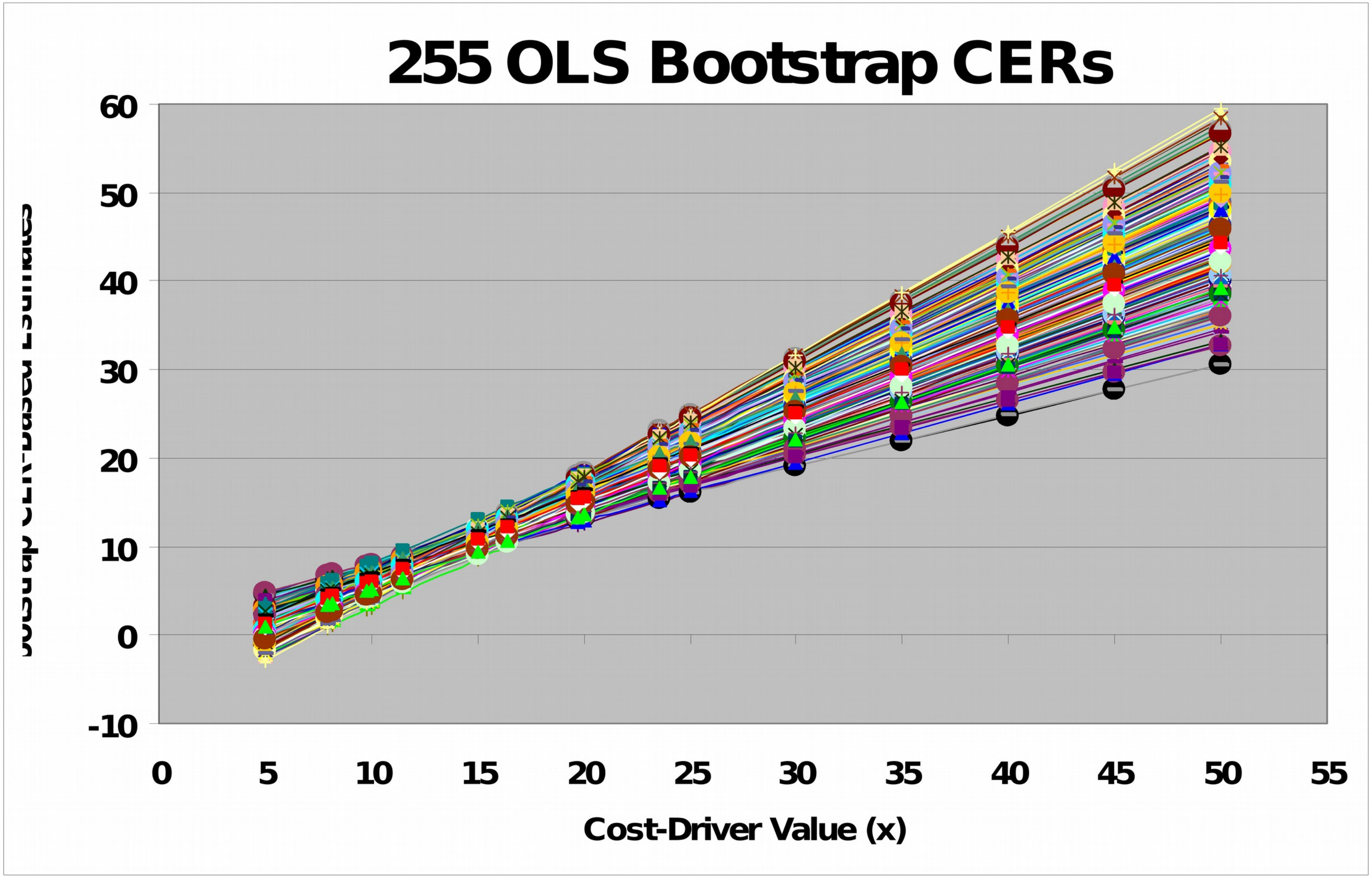
Sampling (of the residuals) is done with replacement

x Values (Cost Driver)	Bootstrap Actual = Estimate + Residual	#1	#2	#3	#4	#5	#6	...
7.9	3.699+1st Residual	3.804	5.804	6.034	0.166	0.624	6.118	...
8.2	4.005+2nd Residual	3.651	6.424	4.109	3.651	6.340	6.424	...
9.8	5.635+3rd Residual	5.739	2.102	2.102	7.740	7.970	5.281	...
11.5	7.367+4th Residual	9.785	3.833	9.785	9.785	9.472	9.785	...
16.4	12.358+5th Residual	14.463	14.463	8.825	14.463	12.004	12.004	...
19.7	15.720+6th Residual	17.825	12.187	17.825	17.825	18.139	12.644	...
23.6	19.693+7th Residual	16.159	22.112	19.339	19.797	22.112	19.797	...

Note: Each bootstrap sample is treated as if it were a set of "actual" data.
The only use made of the <u>real</u> actual data set is to calculate the estimates and residuals.

MCR: Graphs of 255 Bootstrap “CERs”

Use each Bootstrap sample to calculate an OLS CER:



TECOLOTE RESEARCH, INC.

MCR: Rank All 255 Estimates Associated with Each Cost-Driver Value X

Estimate Ranks	x = 5	x = 15	x = 50	Estimate Ranks	x = 5	x = 15	x = 50
1	-3.021	8.583	30.604	226	2.861	11.970	52.659
2	-2.434	8.586	30.629	227	2.942	11.971	52.908
3	-2.387	8.805	32.750	228	2.951	11.975	52.953
4	-2.381	8.918	32.791	229	2.989	11.977	53.272
5	-2.378	9.057	32.913	230	2.989	12.009	53.285
6	-2.311	9.096	32.922	231	3.008	12.014	53.289
7	-2.112	9.108	33.173	232	3.012	12.069	53.303
8	-2.100	9.273	34.280	233	3.042	12.073	53.346
9	-1.889	9.301	34.672	234	3.045	12.080	53.437
10	-1.889	9.352	34.862	235	3.205	12.084	53.455
11	-1.848	9.390	35.151	236	3.238	12.098	53.493
12	-1.717	9.426	35.517	237	3.254	12.103	53.535
13	-1.681	9.534	36.022	238	3.289	12.110	53.752
14	-1.680	9.581	36.132	239	3.350	12.114	54.388
15	-1.608	9.582	36.417	240	3.415	12.119	54.509
16	-1.562	9.589	36.497	241	3.439	12.136	54.650
17	-1.526	9.652	36.763	242	3.577	12.146	54.852
18	-1.489	9.655	36.819	243	3.598	12.171	55.028
19	-1.376	9.677	36.843	244	3.890	12.202	55.142
20	-1.338	9.715	36.857	245	3.909	12.209	55.163
21	-1.334	9.728	37.228	246	4.006	12.238	56.649
22	-1.330	9.731	37.388	247	4.161	12.270	56.747
23	-1.293	9.750	37.731	248	4.173	12.305	56.915
24	-1.281	9.762	37.754	249	4.174	12.336	57.034
25	-1.272	9.780	37.848	250	4.274	12.480	57.600
26	-1.235	9.787	38.052	251	4.461	12.534	57.894
27	-1.206	9.796	38.122	252	4.520	12.539	58.082
28	-1.192	9.813	38.444	253	4.556	12.613	58.398
29	-1.144	9.822	38.449	254	4.620	12.659	58.901
30	-1.101	9.851	38.471	255	4.896	13.155	59.453

Upper 10^{th} Percentile

Lower 10^{th} Percentile

MCR: Can We Derive Prediction Bounds from Bootstrap Bounds?

- **It Turns Out that the "Bootstrap Bounds," namely the 10th and 90th Percentile Bootstrap Estimates, for any Given x Value are Closer Together than the Known 80% Lower and Upper OLS Prediction Bounds**
- **After some Theoretical Investigations and Numerical Experimentation, We Found that Adjusting the OLS Bootstrap Bounds Outward by an Additive Amount Equal to the SEE of the "Real" CER Brought the so-called "Bootstrap-Based" Bounds Closer to the OLS Prediction Bounds**

Downside of Using the Bootstrap - PI (1/4)

- **The Bootstrap sampling may not provide representative samples of the error distribution for small samples (see slide #7)**
- **The Bootstrap-based prediction intervals (between 10th and 90th percentiles) are not centered at the regression equation**
  - Not possible to model symmetric distributions using the Bootstrap-based (B-B) PIs in risk analysis
  - Hard to specify the B-B PIs using current risk packages
- **The Bootstrap-based PI bounds (for linear and factor equations) appear to have systematic patterns**

TECOLOTE RESEARCH, INC.

Downside of Using the Bootstrap – PI (OLS CER) (2/4)

Y	Cost Driver X	BB 10% Low	BB 90% High	OLS 10% Low	OLS 90% high	CER Predicted	Avg(L,H)	Avg <>CER	BB_PI <> OLS_PI
	5	-4.003	5.739	-4.214	5.704	0.745	0.868	16.5%	-1.8%
3.595	7.9	-0.574	8.132	-0.931	8.330	3.699	3.779	2.2%	-6.0%
1.900	8.2	-0.204	8.408	-0.598	8.608	4.005	4.102	2.4%	-6.5%
3.300	9.8	1.608	9.822	1.158	10.112	5.635	5.715	1.4%	-8.3%
	10	1.819	9.973	1.375	10.303	5.839	5.896	1.0%	-8.7%
10.900	11.5	3.472	11.336	2.980	11.753	7.367	7.404	0.5%	-10.4%
	15	7.033	14.743	6.582	15.282	10.932	10.888	-0.4%	-11.4%
15.434	16.4	8.360	16.266	7.965	16.751	12.358	12.313	-0.4%	-10.0%
16.074	19.7	11.045	20.108	11.103	20.337	15.720	15.577	-0.9%	-1.8%
	20	11.287	20.446	11.380	20.671	16.025	15.867	-1.0%	-1.4%
17.274	23.6	14.302	24.508	14.617	24.758	19.693	19.405	-1.5%	0.6%
	25	15.488	26.018	15.837	26.401	21.119	20.753	-1.7%	-0.3%
	35	23.347	37.898	24.128	38.483	31.306	30.623	-2.2%	1.4%
	40	27.289	43.897	28.104	44.695	36.399	35.593	-2.2%	0.1%
	45	31.195	49.898	32.018	50.968	41.492	40.547	-2.3%	-1.3%
	50	35.200	56.029	35.890	57.282	46.586	45.615	-2.1%	-2.6%

- The Bootstrap-based PIs (90th - 10th percentile) appear to be narrower than the corresponding OLS PIs
- The mid points of the Bootstrap-based PIs seem to be over-estimating the mean at the low end and under-estimating the mean at the high end

Tecolote Research, Inc.

Downside of Using the Bootstrap – PI (Y=0.6372X) (3/4)

Y	Cost Driver X	BB 10% Low	BB 90% High	MUPE 10% Low	MUPE 90% high	CER Predicted	Avg(L,H)	Avg <> CER	BB_PI <> MUPE_PI
	5	1.022	5.263	0.925	5.446	3.186	3.143	-1.4%	-6.2%
3.595	7.9	1.615	8.315	1.462	8.605	5.034	4.965	-1.4%	-6.2%
1.900	8.2	1.676	8.631	1.518	8.932	5.225	5.154	-1.4%	-6.2%
3.300	9.8	2.004	10.315	1.814	10.675	6.244	6.160	-1.4%	-6.2%
	10	2.045	10.525	1.851	10.893	6.372	6.285	-1.4%	-6.2%
10.900	11.5	2.352	12.104	2.129	12.527	7.328	7.228	-1.4%	-6.2%
	15	3.067	15.788	2.777	16.339	9.558	9.428	-1.4%	-6.2%
15.434	16.4	3.354	17.261	3.036	17.864	10.450	10.308	-1.4%	-6.2%
16.074	19.7	4.028	20.735	3.646	21.459	12.553	12.382	-1.4%	-6.2%
	20	4.090	21.051	3.702	21.786	12.744	12.571	-1.4%	-6.2%
17.274	23.6	4.826	24.839	4.368	25.707	15.038	14.833	-1.4%	-6.2%
	30	6.135	31.575	5.553	32.678	19.116	18.855	-1.4%	-6.2%
	35	7.157	36.838	6.478	38.125	22.302	21.998	-1.4%	-6.2%
	40	8.179	42.101	7.404	43.571	25.488	25.140	-1.4%	-6.2%
	45	9.201	47.364	8.329	49.018	28.674	28.283	-1.4%	-6.2%
	50	10.224	52.626	9.255	54.464	31.860	31.425	-1.4%	-6.2%

- The Bootstrap-based PIs (90th - 10th percentile) are 6.2% narrower than the MUPE PIs for the entire data range, even beyond the data range
- The mid points of the Bootstrap-based PIs are consistently less than the regression line by 1.4%
- This factor CER (see Ref 2), fitted using the ZMPE method, is the same as the MUPE CER

Downside of Using the Bootstrap - PI (4/4)

- **The difference of the upper bounds between OLS Prediction Interval and OLS Confidence Interval for the Mean (UBPI - UBCI) is related to**
  - standard error of the regression (SEE)
  - sample size (n)
  - level of confidence (80%, 90%, etc.)
  - location of the driver variable (x_e) at the estimating point, and
  - sample standard deviation of the independent variable

Hence the difference, UBPI - UBCI, does not equal to SEE

$$(t_{\alpha/2,df}) * SEE * \left(\sqrt{1 + \frac{1}{n} + \frac{(x_e - \bar{x})^2}{\sum (x_i - \bar{x})^2}} - \sqrt{\frac{1}{n} + \frac{(x_e - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right) \neq SEE$$

- **What adjustments should we choose for different confidence levels, e.g., 15th vs. 85th, or 5th vs. 80th?**
- ***Further research on the "Bootstrap Bound ± SEE" method to derive prediction intervals from CERs of general form is warranted*** **(Ref 1)**

Difficulties Noted in Ref (1)

- **It was noted in Reference 1 that Variance of the Estimate (i.e., $\hat{y}$) for General Nonlinear CERs (as well as IRLS CERs) has not been solved**
  - Reference 1 stated that the Prediction Interval Problem Appears Not to Have Been Solved; Research in this Direction May be Worthwhile
  - While We Await the “Exact” Theoretical Solution to be Found (which, if history is a guide, could take decades), It Would be Useful to Have Available a Practical “Ad Hoc” Method that We Can Apply to Generate Prediction Intervals in Any Particular Case
- **It is worthwhile to do a literature search on PI for general nonlinear equations**

Literature Search on PI for Nonlinear Equations

- **SAS/STAT 9.1 Users Guide contains useful info**
  - http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/stat_ug_7313.pdf
  - Chapter 50 explains the NLIN procedure. SAS/STAT can produce confidence and prediction intervals for nonlinear regression equations.
- **UCLA Academic Technology Services has a SAS library**
  - It contains a section entitled “Nonlinear Regression With The SAS System.” In this section, it shows an example with a nonlinear output using SAS proc code to generate confidence bounds for the mean predictions and prediction intervals for an individual predictions.
- **S-PLUS and the R software code can generate CIs and PIs for nonlinear equations**
- ***Advanced Linear/Non-Linear Models* (add-on product) in Statistica has the PI feature** (confirmed by Jenny Meyer)
- **Systat 11 has an IRLS feature, but does not provide PIs for nonlinear regression**
- **SPSS does not offer PIs for nonlinear regression equations**

Prediction Interval for MUPE Available?

- **The CIs and PIs are included in the nonlinear regression output in SAS when certain options are specified in the output command, e.g., L95M=, U95M=, L95=, and U95=**
- **Since the MUPE CERs are generated by *weighted* least squares in unit space, the regression outputs (including PIs) should be available by running the appropriate command file. For example, all USCM7 and USCM8 CERs were developed by running the tailored script files (command files) in Systat 5**

Derive Solution by Taylor Series Expansion

- **Taylor series using first order derivatives:**

$$\boxed{f(\underline{\mathbf{x}},\underline{\boldsymbol{\theta}}) \cong f(\underline{\mathbf{x}},\underline{\boldsymbol{\theta}}^o) + \left.\frac{\partial f}{\partial \underline{\boldsymbol{\theta}}}\right|_{\underline{\boldsymbol{\theta}}=\underline{\boldsymbol{\theta}}^o} (\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^o)}$$

- **Taylor series using second order derivatives :**

$$\boxed{f(\underline{\mathbf{x}},\underline{\boldsymbol{\theta}}) \cong f(\underline{\mathbf{x}},\underline{\boldsymbol{\theta}}^o) + \left.\frac{\partial f}{\partial \underline{\boldsymbol{\theta}}}\right|_{\underline{\boldsymbol{\theta}}=\underline{\boldsymbol{\theta}}^o} (\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^o) + \tfrac{1}{2}(\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^o)' \left.\frac{\partial^2 f}{\partial \underline{\boldsymbol{\theta}}^2}\right|_{\underline{\boldsymbol{\theta}}=\underline{\boldsymbol{\theta}}^o} (\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^o)}$$

- **Derive the solution using first order derivatives:**
  - $V(\hat{y}|_{x=x_o}) \cong z_o(Z'WZ)^{-1}z_o'\sigma^2$ for weighted least squares
    $Z = \partial f/\partial\theta$ (evaluated at $\hat{\theta}$),
    W = weighting matrix (W = **I** if not weighted)
  - $V(y - \hat{y}|_{x=x_o}) \cong (f(x_o)^2 + z_o(Z'WZ)^{-1}z_o')\sigma^2$ (for MUPE CERs)
  - A possible typo in the PI formulas in SAS/STAT 9.1 Users Guide

PI Closed-Form Formula for Simple Weighted Least Squares (a + bx)

- **PI for WLS:**

$$\hat{y}_0\,(when\ X = x_0) \pm t_{(\alpha/2,\,df)} * Se * \sqrt{\frac{1}{w_o} + \frac{1}{\sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{SS_{wxx}}}$$

where:

$$\bar{x}_w = \sum_{i=1}^{n} w_i\,(x_i) \Big/ \sum_{i=1}^{n} w_i$$

$$SS_{wx} = \sum_{i=1}^{n} w_i\,(x_i - \bar{x}_w)^2$$

Note: w_i is the weighting factor for the ith data point and w_o is the weighting factor for y when $x = x_o$.

- **If the data set is unavailable, we have a heuristic approach to approximate the PIs based upon the formula above**

TECOLOTE RESEARCH, INC.

Develop PIs in Excel for Five Sample Cases

- **A linear model with two drivers: $y = a + bX + cZ + \varepsilon$**
  - The PIs (generated manually in Excel) match exactly the PIs reported in CO$TAT.
- **A Michaelis-Menten model: $y = Vmax / (k + X) + \varepsilon$**
  - CO$TAT nonlinear output (estimates, asymptotic se, LS Summary stats, etc.) matches the SAS output exactly.
  - The PIs (generated manually in Excel) match exactly the PIs reported in SAS.
- **A log-logistic model: $y = \delta + (\alpha - \delta) / (1+\exp(\beta*\log(X/\gamma))) + \varepsilon$**
  - CO$TAT nonlinear output (estimates, asymptotic se, LS Summary stats, etc.) matches the SAS output exactly.
  - The PIs (generated manually in Excel) follows closely the PIs reported in SAS. The largest error (for 97.5th percentiles) is around 0.16%.
- **A weight based MUPE CER: $y = a(W)^b$**
- **A MUPE factor CER: $y = a(X)$**

TECOLOTE RESEARCH, INC.

Conclusions – Analytic Method

- **The analytic method to derive PIs for nonlinear equations is superior to the Bootstrap heuristic approach**
  - References are available for computing the variance of the estimate (i.e., $\hat{y}$) regarding general nonlinear CERs
- **Confidence and prediction intervals for nonlinear CERs are already given in SAS, S-PLUS, R, Statistica, etc.**
- **The PI calculations can be used for MUPE CERs as well because they are generated by weighted least squares**
  - Need to verify a few more cases using SAS (or S-PLUS) for weighted least squares.

Conclusions – The Bootstrap Method Concerns (for PI)

- **The Bootstrap results may not be reliable for small samples**
  - Cost analysts usually deal with small (actual) samples
- **As noted in Reference 1, we might need many Bootstrap samples (such as 10,000) to construct PI bounds**
- **Since nonlinear regression equations may not converge, we might need to generate more Bootstrap sample CERs to overcome the shortage**
- **The nonlinear curve fitting process may be caught in local minimums, which may impact the solution**
- **The Bootstrap-based PI bounds seem to have systematic patterns for linear and factor equations; worth further study**

References

1. Book, S., “Prediction Intervals for CER-Based Estimates (With Cost Driver Values Outside the Data Range),” 37th DoDCAS, Williamsburg, VA, 10-13 February 2004
2. Book, S., “Prediction Intervals for General-Error-Regression CERs),” 39th DoDCAS, Williamsburg, VA, 14-17 February 2006
3. SAS/STAT 9.1 Users Guide.
4. Schabenberger, Oliver, “Nonlinear Regression in SAS,” SAS Library, UCLA Academic Technology Services.
5. http://lib.stat.cmu.edu/R/CRAN/ (StatLib is hosted by the Department of Statistics at Carnegie Mellon University)

TECOLOTE RESEARCH, INC.
Bridging Engineering and Economics
Since 1973

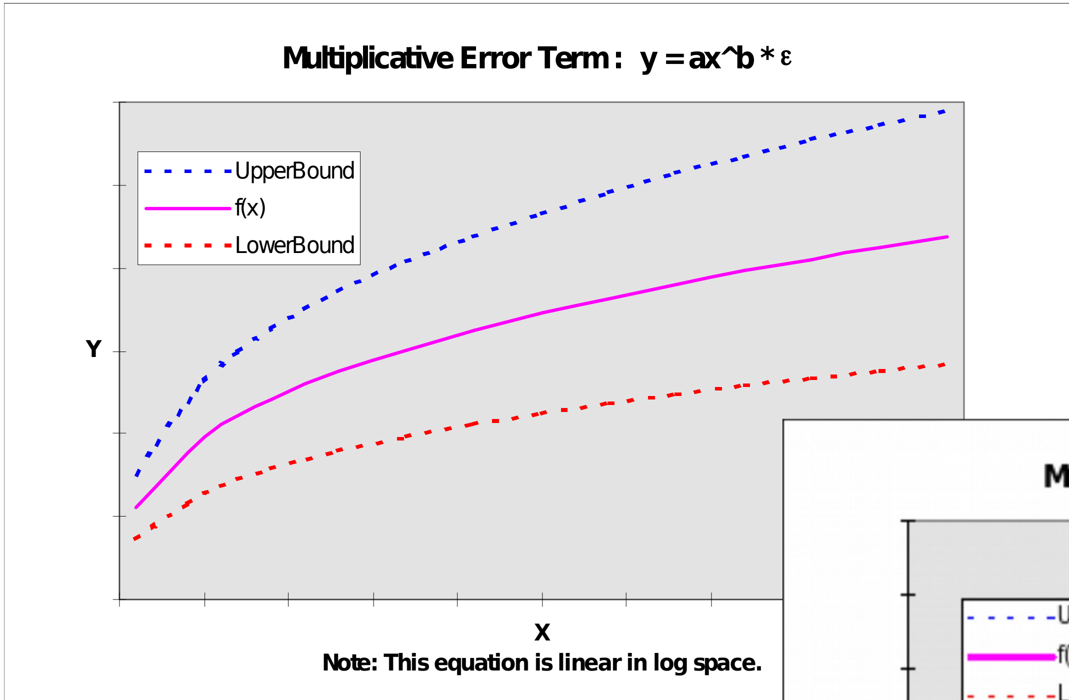
Backup Slides

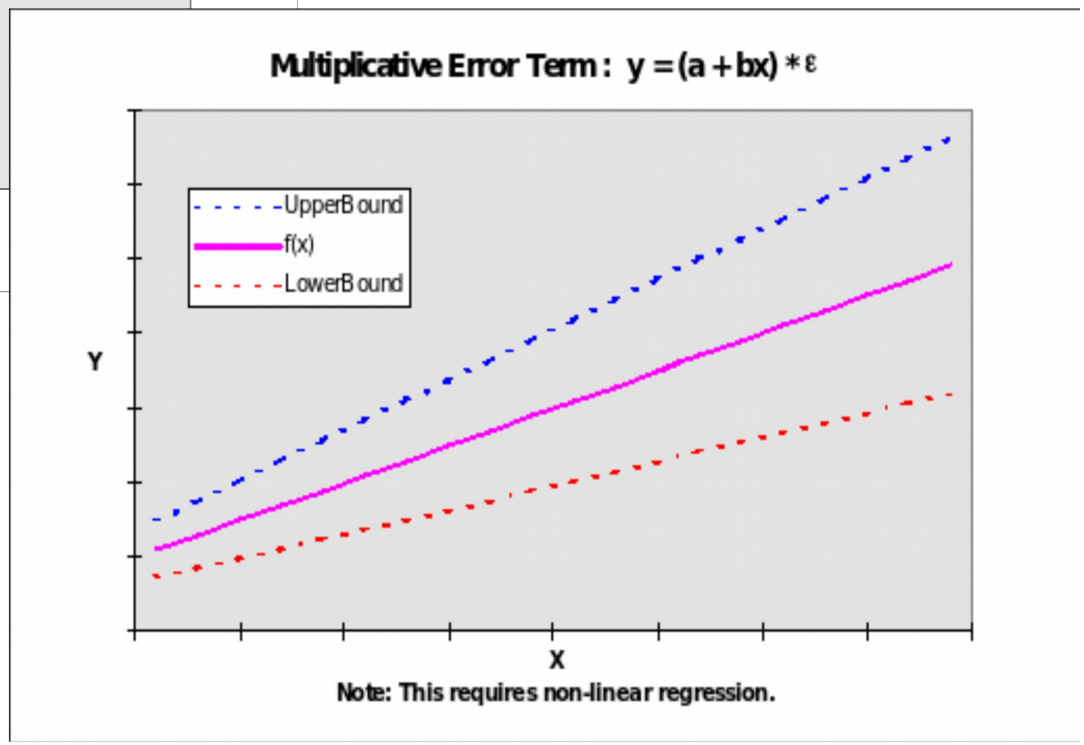
Develop PIs in Excel for Five Sample Cases (2/2)

- **A log-logistic model: y = δ + (α - δ) / (1+exp(β*log(X/γ))) + ε**
  - CO$TAT nonlinear output (estimates, asymptotic se, LS Summary stats, etc.) matches the SAS output exactly.
  - The PIs (generated manually in Excel) follow closely the PIs reported in SAS. The largest error is around 0.16%.

SAS		First derivatives		
L95IND	**U95IND**	**L95**	**U95**	**% Delta**
77.894	116.296	77.960	116.229	-0.06%
77.307	115.357	77.332	115.331	-0.02%
74.683	112.511	74.681	112.513	0.00%
63.095	102.009	63.239	101.865	-0.14%
33.831	72.531	33.819	72.543	0.02%
0.961	39.739	0.995	39.705	-0.09%
-13.694	25.482	-13.735	25.522	0.16%

Multiplicative Error Term





Cost variation is proportional to the scale of the project

Multiplicative Error Model – MUPE

Definition of cost variation for Y = f(x)*ε

- **MUPE: E(ε) = 1, V(ε) = σ² ⇨ Weighted least squares in unit space**
  - Error = (Y-f(x)) / f(x) = (ε – 1)
  - $$Minimize \sum_{i=1}^{n} \left[\frac{y_i - f(x_i)}{f_{k-1}(x_i)} \right]^2$$

Note:

E((Y-f(x)) / f(x)) = 0

V((Y-f(x)) / f(x)) = σ²

where k is the iteration number